Atomic Spatial Processes

Sean Jewell SEAN.JEWELL@STAT.UBC.CA
Neil Spencer NEIL.SPENCER@STAT.UBC.CA
Alexandre Bouchard-Côté BOUCHARD@STAT.UBC.CA
Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

Abstract

The emergence of compact GPS systems and the establishment of open data initiatives has resulted in widespread availability of spatial data for many urban centres. These data can be leveraged to develop data-driven intelligent resource allocation systems for urban issues such as policing, sanitation, and transportation. We employ techniques from Bayesian non-parametric statistics to develop a process which captures a common characteristic of urban spatial datasets. Specifically, our new spatial process framework models events which occur repeatedly at discrete spatial points, the number and locations of which are unknown a priori. We develop a representation of our spatial process which facilitates posterior simulation, resulting in an interpretable and computationally tractable model. The framework's superiority over both empirical grid-based models and Dirichlet process mixture models is demonstrated by fitting, interpreting, and comparing models of graffiti prevalence for both downtown Vancouver and Manhattan.

1. Introduction

There has been a recent rise in the quantity and quality of spatial data due to the emergence of new technologies, such as the widespread availability of cheap and compact global positioning device systems (GPS). The availability of this new data represents an opportunity in machine learning and artificial intelligence to develop intelligent resource allocation systems for urban centres. Accordingly, these opportunities have been accompanied by a rise in the prominence of spatial statistics, in particular urban spatial statistics. Data-driven analyses have been conducted in areas such as transportation (Páez & Scott, 2005), health (Boruff et al., 2012), and crime (Malleson & Andresen, 2014). In particular, applying machine learning to predictive policing (Short et al., 2008; Mohler et al., 2011; Wang et al., 2012; 2013) calls for new types of spatial models.

Spatial datasets are often comprised of geospatial coordinates recording locations of phenomena of interest (e.g. graffiti locations, shown in Figures 1a and 1c). A classical modelling approach involves conceptualizing the data as a realization of a Poisson process (PP). A PP is parametrized by a rate measure from which all observations are independently chosen. Lately, attention has been devoted to Bayesian non-parametric approaches to modelling this rate measure (e.g. Gelfand et al. (2005); Guindani et al. (2009); Taddy (2010); Ding et al. (2012)).

PP models employing continuous (or piecewise continuous) rate measures are not appropriate for all urban spatial datasets, specifically those datasets in which multiple observations can occur in the same location. Henceforth, we refer to such datasets as atomic data. A location exhibiting multiple observations demonstrates positive probability associated with that location. Informally, points with positive probability violate the assumption of a piecewise continuous rate measure, in which all points should be (almost-surely) unique. We formalize this intuition in terms of predictive power below.

In order for a rate measure to capture multiple identical observations, it must possess atoms (points of positive probability) at the location of these observations. We call such rate measures atomic rate measures. Note that in this paper, we employ purely atomic rate measures. That is, we use rate measures with no continuous components. We utilize urban graffiti locations as a running example of atomic data throughout this paper. Consider Figure 1a which illustrates the locations of graffiti in Vancouver. There are many locations possessing more than five incidents of graffiti, drastically violating the assumption of a continuous rate measure. The New York OpenData catalog contains many atomic urban datasets which are relevant to designing resource allocation systems, such as noise complaints, housing code violations, and service requests. Manage-

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

ment of these areas could benefit from the development of intelligent automated research allocation systems.

Atomicity in urban datasets can arise naturally (e.g., one graffiti artist tags over a rival artist's graffiti), or as an artifact of the data collection process. Typically, artificial atomicity arises when spatial observations within a certain area are aggregated to a single point (e.g. nearest civic address, middle of street, nearby landmark, etc.) during the data collection process. This may be due to various reasons, such as to preserve anonymity, simplify data collection, or to represent uncertainty in measurements. In many of these instances, the aggregated (approximate) locations provide the desired information (e.g. the address of an often vandalized building indicates to police officers the location to monitor). In addition, the artificial atoms can reveal additional insights into data (e.g. multiple noise complaints in close proximity refer to a common noise problem), which are not as apparent without atoms. Thus it is often practical, and even beneficial, to directly model the atoms in artificially atomic data. Even in cases where exact locations are preferred, it is often more practical to directly model the artificial atoms. Otherwise, the alternative is to view the exact locations of observations as latent variables. There is often limited (or no) information available regarding the mechanism of data aggregation, making such a modelling approach difficult. Even when information is available, assigning each observation a latent location variable can quickly lead to a model for which inference is computationally intractable.

Appropriately capturing the atomic nature of urban data is vital for predicting future observations. Even if a point has been observed repeatedly in the past (e.g., a graffiti location over which rival artists dispute), a continuous rate measure would assign zero probability to that location reoccurring, whereas atomicity can assign positive probability. In addition, accounting for atoms can lead to insight regarding the process underlying a dataset (e.g., graffiti tends to encourage more graffiti on the same exact building). Despite these strengths, we are unaware of existing non-parametric spatial approaches based on atomic rate measures. One potential explanation is the challenge of maintaining joint uncertainty on the cardinality and locations for the set of atoms while still using the topological information and maintaining tractability (in the sense of exact approximate inference algorithms).

We propose a model for atomic spatial data that addresses these challenges. We do this by leveraging tools from the Bayesian non-parametric literature and applying them to a new domain. Henceforth, we refer to our new model as an atomic spatial process (ASP). Specifically, the key features of the ASP can be organized into three distinct levels; a Poisson process (of which the observed data is a realization), a gamma process (GaP), and a Dirichlet process mixture model (DPM). The GaP specifies a distribution over purely atomic measures which acts as our prior for the PP rate measure, and the DPM is the prior on the base measure of the GaP. In the graffiti context, these levels have intuitive meanings, making the model posterior distributions interpretable. The GaP assigns a propensity of being graffitied to all possible graffiti locations (e.g., buildings), and the DPM represents the distribution of unique locations of graffiti (with each mixture component being a distinct graffiti hotspot).

Note that individual components of the ASP framework have been previously applied in spatial problems. For example, Wolpert & Ickstadt (1998) use gamma processes as a component of their model, but their framework also involves a smoothing component which ultimately results in a continuous rate measure. It is worth noting that, despite what their name may suggest, spatially normalized gamma processes (Rao & Teh, 2009) are not specialized gamma processes for spatial applications. Instead, "spatially" refers to an unrelated notion of defining a gamma process over an augmented general "space" and subsequently deriving dependent random measures by collapsing and normalizing portions of this gamma process. Other than the use of normalized gamma processes, this technique is unrelated to the ASP framework.

The remainder of this paper is organized as follows. Section 2 consists of a brief review of some Bayesian nonparametric modelling concepts. In Section 3, we formalize ASPs and develop a representation which facilitates posterior simulation. In Section 4, we provide an approximate posterior inference algorithm for ASPs. In Section 5, we demonstrate the performance and interpretability of ASPs using two graffiti datasets. Section 6 illustrates the benefits gained by using ASP to model atomic data by comparing its performance with a Dirichlet process mixture model, an empirical model, and a mixture of the two. Section 7 provides some final remarks.

2. Background and Notation

Bayesian non-parametric modelling involves the use of Bayesian models with infinite-dimensional parameter spaces—typically, the marginals of a stochastic process—which provide enough flexibility to capture complex relationships in data for which limited knowledge of structure is known a priori. Bayesian non-parametric constructions also allow the model to dynamically evolve as more data become available. See Hjort et al. (2010) for a review of Bayesian non-parametric modelling.

The Poisson process is a popular choice for modelling random collections (Kingman, 1992, Sec. 1-2) of points over

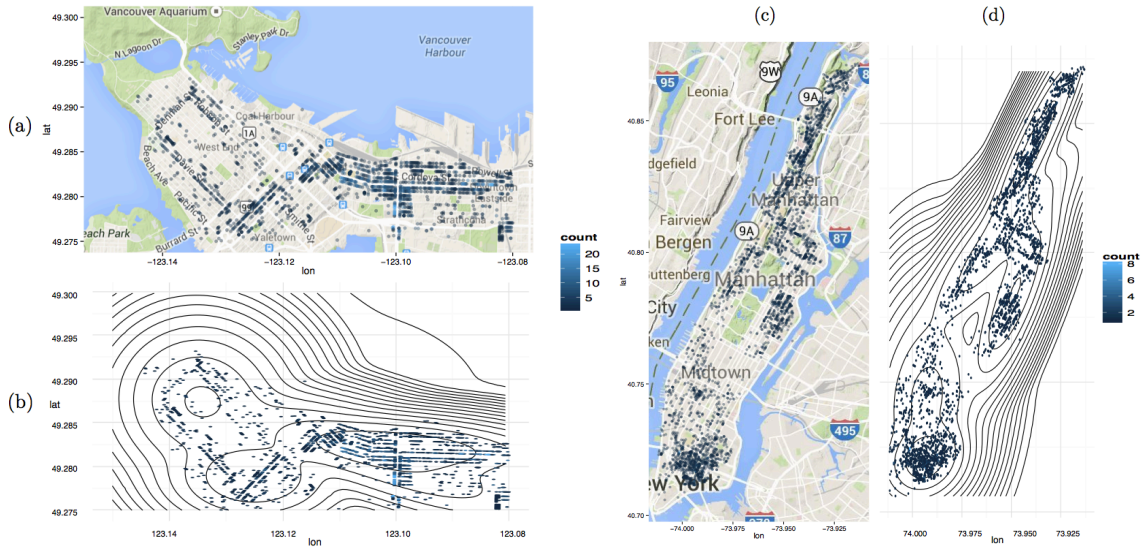

Figure 1. a) A visualization of 7675 graffiti in 1982 unique locations of downtown Vancouver. b) Visualization of the predictive distribution for the next piece of graffiti in downtown Vancouver obtained using an Atomic Spatial Process model . c) A visualization of 3946 graffiti in 3406 unique locations of Manhattan. d) Visualization of the predictive distribution for the next piece of graffiti in Manhattan obtained using an Atomic Spatial Process model.

a space Ω (in our applications, $\Omega \subset \mathbb{R}^2$, and Ω represents latitudes and longitudes over a given area). These collections typically represent locations of events within a specified time frame: for example, locations of crime scenes in a city over a given year. Poisson processes are parameterized by a rate measure μ on Ω which specifies the propensity of events on each region in Ω. In spatial applications, it is reasonable to assume that μ has a finite normalization, $\mu(\Omega) < \infty$. Given a rate measure μ, generating a realization from the corresponding Poisson process can be done using the following two-step algorithm. First, one samples $N \sim \text{Pois}(\mu(\Omega))$, where $\text{Pois}(\lambda)$ denotes a Poisson distribution with mean λ. This specifies the number of points in the realization. The second step is to generate $X_1, \ldots, X_N \overset{iid}{\sim} \bar{\mu}$ where $\bar{\mu}$ denotes $\mu/\mu(\Omega)$, a probability distribution proportional to μ. Finally, the multiset $\{X_1, \ldots, X_N\}$, denoted X, gives us a realization of the Poisson process ($\text{PP}(\mu)$).

We will also make use of the gamma process (GaP). A realization of a GaP can be viewed as a countably infinite number of triplets $(x_{1,1}, x_{2,1}, w_1), (x_{1,2}, x_{2,2}, w_2), \ldots$ with each $(x_{1,i}, x_{2,i}) \in \Omega$ representing a location (latitude, longitude) and $w_i \in \mathbb{R}^+$ specifying a weight for that location. The generative process of this GaP consists of sampling from a $\text{PP}(\nu)$ with a rate measure $\nu : \Omega \times \mathbb{R}^+ \to \mathbb{R}^+$ given by $\nu(\mathrm{d}x_1, \mathrm{d}x_2, \mathrm{d}w) = w^{-1}e^{-cw}\alpha_0 G_0(\mathrm{d}x_1, \mathrm{d}x_2)$, where ν is the product of a *base measure* $\alpha_0 G_0$ (a parameter of the GaP composed of a probability measure G_0 scaled by a positive constant α_0) with an improper gamma distribution with concentration parameter c. This choice of ν ensures that the sum of all the weights $\sum_{i=1}^{\infty} w_i$ is finite with $\mathbb{E}(\sum_{i=1}^{\infty} w_i) = c^{-1}\alpha_0$ (by Campbell's Theorem).

A realization of a GaP(c^{-1}, α_0, G_0) can be expressed as

$$G = \sum_{i=1}^{\infty} w_i \delta_{(x_{1,i}, x_{2,i})}, \tag{1}$$

where δ indicates the Dirac-delta distribution. Normalizing G given in Equation (1) by replacing each w_i with $p_i = w_i/(\sum_{j=1}^{\infty} w_j)$ results in a random probability measure called a Dirichlet Process (denoted DP(α_0, G_0)) (Ferguson, 1973; Jordan, 2010). Much like a GaP(c^{-1}, α_0, G_0) can be normalized to a DP(α_0, G_0), a DP(α_0, G_0) can be scaled to a GaP(c^{-1}, α_0, G_0) by replacing each p_i with $w_i = \Gamma p_i$ where $\Gamma \sim \text{Gamma}(\alpha_0, c^{-1})$ (shape parameter α_0 and inverse scale c). Dirichlet processes are often used in conjunction with a parametric family indexed by realizations of the DP. Each data point is then obtained in a two-step process: first, sampling a parametric family index or parameter θ from the DP, and second, sampling from the member of the parametric family given θ. This construction is known as a Dirichlet process mixture model (Antoniak, 1974), denoted DPM.

3. Atomic Spatial Processes

Suppose we observe a multiset realization $\{X_1, \ldots, X_N\}$ of a random process over a space Ω. Our goal is to model this process under the assumption that some locations may occur several times in X, and that the total number of observations N is random.

Generative process: An ASP is a Bayesian nonparametric model composed of three distinct levels. We start with a bird's-eye-view of the model, and then describe and motivate each level in detail. At a high-level, the ASP model is structured as follows:

$$\begin{aligned} X|\mu &\sim \mathrm{PP}\left(\mu\right) \\ \mu|G_0 &\sim \mathrm{GaP}\left(c^{-1}, \alpha_0, G_0\right) \\ G_0 &\sim \mathrm{DPM}\left(\alpha_0^{\pi}, H_0\right), \end{aligned}$$

where $H_0(\cdot|\kappa, \nu, \vartheta, \Delta)$ is a bivariate Normal-Inverse-Wishart measure (Gelman et al., 2013, p. 87-88) with ν degrees of freedom, scale parameter κ, mean parameter ϑ, and covariance parameter Δ. The DPM is based on the bivariate Normal family parameterized by a mean and a covariance matrix. We hold $\nu, \kappa, \vartheta, \Delta, c, \alpha_0$ and α_0^{π} fixed for now, but discuss putting priors on the univariate real parameters at the end of the next section.

We start by motivating the sampling process for μ. As reviewed in the previous section, when μ is a continuous measure, all observations in X are distinct with probability one. To avoid this restriction, we therefore utilize a Poisson process with a discrete rate measure, by assuming the measure μ is a realization of a gamma process (GaP) over $\Omega \subset \mathbb{R}^2$ (latitudes and longitudes). However, in a finite observed sample, we do not observe all possible locations in the population. The use of a GaP prior on μ captures this property through the infinite number of atoms in a GaP realization. Any observed data come from a finite number of these atoms, resulting in unobserved atoms that should be accounted for when performing prediction.

Another common feature in urban data is the existence of several "hotspots"—high activity areas, such as neighbourhoods, consisting of many unique locations close in proximity. Such hotspots can be observed in Figures 1a and 1c. Therefore, it is sensible to incorporate proximity to hotspots when assigning predictive density for unobserved atom locations. For this effect, a DP is used in the development of G_0, the base measure for the GaP(c^{-1}, α_0, G_0) prior placed on the rate measure μ. Specifically, we assume G_0 is a Dirichlet Process mixture model (Antoniak, 1974) where the mixture components are bivariate Normal distributions. The following provides an explicit recipe for G_0.

Consider a Dirichlet Process DP(α_0^{π}, H_0) with H_0 defined on the space of parameters of the bivariate Normal family, $\mathbb{R}^2 \times \mathbb{R}^{2\times 2}$. We employ a Normal-Inverse-Wishart measure for H_0, allowing us to exploit the conjugacy between the Normal and Normal-Inverse-Wishart distributions. Let $(\xi_1, \Sigma_1, \pi_1), (\xi_2, \Sigma_2, \pi_2), \ldots$ denote a realization from this process, with $\xi_i \in \Omega$, $\Sigma_i \in \mathbb{R}^{2\times 2}$, and $\pi_i \in [0, 1]$. We build the mixture G_0 by letting $G_0(\cdot) = \sum_{i=1}^{\infty} \pi_i \mathcal{N}(\cdot|\xi_i, \Sigma_i)$, where $\mathcal{N}(\cdot|\xi, \Sigma)$ denotes the bivariate Normal measure with mean ξ and covariance matrix Σ.

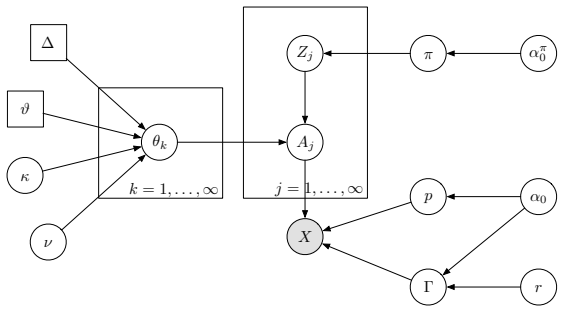

Figure 2. Graphical model formulation of ASP. The random measure μ is encapsulated by the nodes A, Γ, p, and the random measure G_0 is encapsulated by the nodes θ and π. Details for this representation are included in the body of the text. The variables Δ, ϑ, κ, and ν denote the hyperparameters of the Normal-Inverse-Wishart measure H_0. The vector Z summarizes the mixture components in G_0 from which the atoms in μ are drawn. The node X represents a vector consisting of the observations.

This choice of G_0 implicitly assumes that the locations in the modelled data are drawn from a multi-modal distribution over Ω, where the number of modes is unknown. This assumption captures the hotspot structure of many urban data sets; each mode represents a high activity area. Next, we illustrate how each of these processes fit together within our modelling framework.

Collapsed representation: Inference on the model in the form described in the previous section is challenging due to the countably infinite list of continuous latent variables. We now show that this issue can be avoided by analytically marginalizing (collapsing) these infinite lists. The result is a collapsed representation depending solely on the real-valued hyperparameters and the latent cluster memberships Z_j. We start by reformulating the process into the graphical model given by Figure 2. This formulation still involves infinite objects, but uses more explicit random variables to represent the random measures μ and G_0. We will then perform the marginalization on this expression. We now elaborate on how the representation of Figure 2 is defined.

The GaP μ is represented by countably infinite random vectors A (with realization a) and p, as well as the real-valued random variable Γ. The vector $A = (A_1, A_2, \ldots)$ is composed of locations for atoms in μ, and the vector $p = (p_1, p_2, \ldots)$ is composed of the normalized weights assigned to these atoms. The random variable Γ scales p such that its sum is $\mu(\Omega)$. Consequently, the infinite list of atoms $(A_1, \Gamma p_1), (A_2, \Gamma p_2), \ldots$, etc. characterizes μ (by the gamma-DP rescaling property in Section 2).

The DPM $G_0(\alpha_0^{\pi}, H_0)$ is represented using a standard stick-breaking construction (Sethuraman, 1991), denoted by the infinite random vectors θ and π. The vector $\theta = (\theta_1, \theta_2, \ldots)$ is composed such that each θ_k is the param-

eter set (mean and covariance matrix) of the k^{th} Normal mixture component in G_0. The vector $\pi = (\pi_1, \pi_2, \ldots)$ is composed of the weights assigned to these mixture components. The standard strategy for simulating from a mixture model is composed of two steps: first sample a mixture component index then sample from that component. Our formulation follows this process, with the variable Z indicating the mixture component of G_0 corresponding to each A_j. We also define the beta-stick breaking weights β_k^π associated with vector π through the recursive definition $\beta_k^\pi = \pi_k / \prod_{i=1}^{k-1}(1-\beta_i^\pi)$ with $\beta_1^\pi = \pi_1$. Similarly, β_k^p contains the stick-breaking weights associated with p.

We now illustrate how to collapse the model to a more tractable form. The joint distribution of the graphical model in Figure 2 is given by: $\mathbb{P}(\mathrm{d}\beta^\pi, \mathrm{d}\beta^p, \mathrm{d}\theta, \mathrm{d}z, \mathrm{d}a, \mathrm{d}x, \mathrm{d}\gamma, N = n) =$

$$\left[\prod_{k=1}^{\infty} \mathrm{Beta}\left(\mathrm{d}\beta_k^\pi | 1, \alpha_0^\pi\right) H_0(\mathrm{d}\theta_k)\right] \times \left[\prod_{j=1}^{\infty} \mathrm{Mult}\left(\mathrm{d}z_j | \pi\right) L(\mathrm{d}a_j | \theta_{z_j})\right] \times \mathrm{Gamma}\left(\mathrm{d}\gamma | \alpha_0, c^{-1}\right) \left[\prod_{j=1}^{\infty} \mathrm{Beta}\left(\mathrm{d}\beta_j^p | 1, \alpha_0\right)\right] \times \mathrm{Pois}(n|\gamma) \left[\prod_{i=1}^{n} p_{j(a,x_i)} \delta_{x_i \in a}(\mathrm{d}x)\right],$$

where $j(a, x_i)$ denotes the almost sure unique distinct location index j such that $a_j = x_i$, $\mathrm{Beta}(\cdot|\alpha, \beta)$ denotes a beta distribution, $\mathrm{Mult}(\cdot|\pi)$ denotes the multinomial distribution for a single draw given a vector of probabilities π, and $L(\cdot|\theta)$ is a bivariate Normal likelihood. Define $\rho(\cdot)$ to be a function which takes as input an infinite list, and returns a partition (a set of sets of integers) which groups the list indices by the value of their entries. In particular, $\rho(z)$ denotes a partition of the atom indices by their corresponding mixture components. Let a_B denote those atoms whose indices lie in $B \in \rho(z)$. Using conjugacy we can integrate over θ:
$\mathbb{P}(\mathrm{d}\beta^\pi, \mathrm{d}\beta^p, \mathrm{d}z, \mathrm{d}a, \mathrm{d}x, \mathrm{d}\gamma, N = n) =$

$$\left[\prod_{k=1}^{\infty} \mathrm{Beta}\left(\mathrm{d}\beta_k^\pi | 1, \alpha_0^\pi\right)\right] \left[\prod_{j=1}^{\infty} \mathrm{Mult}\left(\mathrm{d}z_j | \pi\right)\right] \times \mathrm{Gamma}\left(\mathrm{d}\gamma | \alpha_0, c^{-1}\right) \left[\prod_{j=1}^{\infty} \mathrm{Beta}\left(\mathrm{d}\beta_j^p | 1, \alpha_0\right)\right] \times \mathrm{Pois}(n|\gamma) \left[\prod_{i=1}^{n} p_{j(a,x_i)} \delta_{x_i \in a}(\mathrm{d}x)\right] \times \prod_{B \in \rho(z)} \int H_0(\mathrm{d}\theta) \prod_{i \in B} L(\mathrm{d}a_i | \theta).$$

Defining the marginal distribution of the parametric model $m(\cdot)$ (which can be computed by conjugacy), as $m(\mathrm{d}a) = \int H_0(\mathrm{d}\theta) L(\mathrm{d}a|\theta)$, our expression simplifies to[1]: $\mathbb{P}(\mathrm{d}\beta^\pi, \mathrm{d}\beta^p, \mathrm{d}z, \mathrm{d}a, \mathrm{d}x, N = n) =$

$$\left[\prod_{k=1}^{\infty} \mathrm{Beta}\left(\mathrm{d}\beta_k^\pi | 1, \alpha_0^\pi\right)\right] \left[\prod_{j=1}^{\infty} \mathrm{Mult}\left(\mathrm{d}z_j | \pi\right)\right] \times \left[\prod_{B \in \rho(z)} m(\mathrm{d}a_B)\right] \left[\prod_{j=1}^{\infty} \mathrm{Beta}\left(\mathrm{d}\beta_j^p | 1, \alpha_0\right)\right] \times \left[\prod_{i=1}^{n} p_{j(a,x_i)} \delta_{x_i \in a}(\mathrm{d}x)\right] \mathrm{NegBin}\left(n \middle| \frac{1}{c+1}, \alpha_0\right),$$

where $\mathrm{NegBin}(\cdot|p, r)$ denotes an extended negative binomial distribution with probability of success p and (real valued) failure count r. The negative binomial distribution arises from marginalizing a gamma distributed prior on a Poisson distribution rate parameter. Marginalization over both β^π, β^p leads to our final expression, namely an explicit and tractable expression for

$$\mathbb{P}(\mathrm{d}z, \mathrm{d}a, \mathrm{d}x, N = n) = \left(\mathrm{CRP}\left(\rho(\mathrm{d}z) | \alpha_0^\pi\right) \prod_{B \in \rho(z)} m(\mathrm{d}a_B)\right) \times \mathrm{NegBin}\left(n \middle| \frac{1}{c+1}, \alpha_0\right) \mathrm{CRP}\left(\rho(\mathrm{d}x) | \alpha_0, N = n\right), \tag{2}$$

where $\mathrm{CRP}(\cdot|\alpha)$ denotes the Chinese Restaurant process table assignment probability measure (Aldous, 1985) with parameter α_0, a function which is easy to evaluate.

For a given set of hyperparameters, this collapsed likelihood is a function of $\rho(z)$, the mixture component memberships of the atoms in μ, and $\rho(x)$, the atom assignments of the observations. By the construction of μ, all atom locations are unique with probability 1. Therefore, the locations of the observations in x completely specify $\rho(x)$, leaving $\rho(z)$ as the only combinatorial random variable. Unfortunately, the number of possibilities of $\rho(z)$ is intractably large for most applications. Additionally, it is often beneficial to resample some of the hyperparameters. We propose a Markov chain Monte Carlo algorithm to conduct approximate inference of the posterior in Section 4.

The marginalization of continuous variables results in a mixed posterior rate measure μ. That is, the posterior distribution of μ is composed of discrete point masses at observed graffiti points as well as density at all other points in space. The continuous part is due to the unobserved atoms discussed when motivating G_0 in Section 3. As was

[1] To streamline presentation, we use the following abuse of notation. For a random variable τ, $\mathrm{d}\tau$ denotes, depending on context, either a subset T in the range of τ, or the event $(\tau \in T)$. This allows us to concisely represent measures and avoid using densities, which would degenerate because of the infinite products.

desired, the density weights are assigned according to the posterior of the DPM G_0, which contains hotspots.

Finally, we considered model extensions placing an additional level of hyper-priors on the parameters $\alpha_0, \alpha_0^\pi, c, \kappa$ and ν. To satisfy the domain of the hyperparameters, we use (shifted) exponential priors with mean 10^{100} for each of the parameters ν, κ, α_0, c, and α_0^π. For example, the domain of ν satisfies $1 < \nu < \infty$, so $\nu - 1 \sim \exp(10^{-100})$. We chose 10^{100} as the mean of the priors on the hyperparameters because a high mean on the exponential distribution makes these priors uninformative. Rerunning the simulations with means of 10^{10} or 10^{1000} yields similar results.

4. Approximate Posterior Inference

We now propose a Markov chain Monte Carlo (MCMC) strategy for sampling the cluster membership variables and the hyperparameters of the model. The state space of the largest model considered contains the clustering z and five real numbers ($\alpha_0, \alpha_0^\pi, c, \kappa$ and ν). The 2×2 covariance matrix Δ and the mean vector ϑ are not resampled; they are fixed at the identity matrix and 0 vector, respectively.

Each sweep of the MCMC algorithm consists of two steps. First, a Gibbs sampling strategy is used to propose a new mixture component membership z_j for each observed unique location a_j. This consists of generating a random permutation P of the locations, followed by sequentially sampling (in order of P) a new value for each z_j according to the conditional distribution obtained from Equation (2). The mixture component memberships are initially configured such that each atom has its own component.

The second step of the sweep consists of performing Metropolis-Hastings moves for select hyperparameters of the model in a randomly permuted order. Starting values for the hyperparameters can be chosen by considering the mixing behaviour for varying short runs. We use random walk Gaussian kernels as the proposal for each Metropolis-Hastings step. Each kernel is centered at the current value of the selected hyperparameter with a fixed variance specific to each hyperparameter. Short runs are used to calibrate the variance of each hyperparameter to ensure proper mixing. The collapsed formulation, Equation (2), allows for tractable acceptance ratio computations.

For each Monte Carlo sample, the predictive density for a new point Y is given by the expression

$$\bar{\mu}'(\,\mathrm{d}y) = \frac{1}{N+\alpha_0}\sum_{i=1}^{N}\delta_{X_i}(\,\mathrm{d}y) + \frac{\alpha_0}{N+\alpha_0}\frac{1}{T+\alpha_0^\pi}\left(\sum_{i=1}^{K}|B_i|m(\,\mathrm{d}y|a_{B_i}) + \alpha_0^\pi m(\,\mathrm{d}y)\right),$$

where K denotes the number of DPM components from which we have observed atoms where $a_{B_i} = \{a_j : z_j = z_i\}$ as before, $|B_i|$ is the number of atoms in mixture component i, and $T = \sum_{i=1}^{K}|B_i|$ is the number of unique locations.

Averaging this predictive density for a run of the sampler provides an estimate of the predictive distribution. We now demonstrate the utility of this predictive distribution for two graffiti datasets.

5. Data Analysis Graffiti Data

This section illustrates two applications of our ASP model and inference strategy using graffiti data from downtown Vancouver and Manhattan. The cities of Vancouver and New York City each maintain records of all graffiti sites identified by city staff. The data is made publicly available through the Vancouver Open Data Catalog and NYC OpenData database.[2] In both datasets, many pieces of graffiti often share identical locations. We illustrate the ASP framework by conducting an analysis on subsets of each of these datasets. Specifically, we use the downtown region for Vancouver and the Manhattan region of New York. We view the set of all recorded graffiti locations for each dataset as a single realization of a spatial PP over the time frame of data collection. The NYC dataset also contains other variables (clean-up status, approximate date and time of creation), but these are not incorporated into our model.

Fitting an ASP can reveal useful results for city officials. The predictive distribution $\bar{\mu}'$ for graffiti occurrences can inform resource allocation for clean-up. As mentioned in Section 3, our model results in a μ' which is a mixture of atoms and a continuous measure G_0', the mean of the posterior for G_0, the base measure of μ. This G_0' provides predictive information for new locations of graffiti, with the α_0 controlling the probability of a new location being chosen. Specifically, this probability is given by $\alpha_0/(N+\alpha_0)$. In addition, the locations of the modes of G_0' can be viewed as hotspots of graffiti activity. These graffiti hotspots can be targeted for police patrols as a preventative measure.

Downtown Vancouver: The downtown Vancouver graffiti dataset reports 7675 pieces of graffiti occurring at 1982 unique locations. The graffiti locations and counts are illustrated in Figure 1a. Figure 1b illustrates the predictive distribution given by the ASP. The inference of $\bar{\mu}$, the normalized version of μ, for Figure 1b was based on an MCMC strategy (outlined in Section 4) consisting of 20,000 itera-

[2]Vancouver Open Data Catalog: `http://data.vancouver.ca/datacatalogue/graffitiSites.htm`. NYC OpenData database: `https://data.cityofnewyork.us/City-Government/DSNY-Graffiti-Information/gpwd-npar`

tions, with resampling of the hyperparameters α_0, α_0^π, c, κ, and ν. The run details, including MCMC trace plots illustrating well-mixing chains are provided in the supplemental materials. The plot is a combination of two components, the discrete atoms (A_j's) located at the 1982 already observed graffiti locations, and the log density of the normal mixture model G_0 indicating the density assigned to new graffiti locations. We also note that through forward simulation from our priors we have found that the Normal-Inverse-Wishart measure for the DPM base measure is able to capture street-like structure.

The mean of the posterior for α_0 is 867.6, indicating that the predictive probability of a new location is 0.102. Consequently, 10.2 percent of the predictive probability is assigned to G_0 and the remaining probability is assigned to already observed graffiti locations (proportionally to the observed counts). This indicates that the contours in Figure 1b represent 10.2 percent of the probability, with the rest assigned to the atoms proportional to observed counts. There appears to be a major graffiti hotspot East of Burrard between Dunsmuir Street and East Hastings. Outside of this graffiti area, there are two smaller high density areas around the West end of Robson and the Davie/Granville intersection. The density is lower around the coasts and close to Stanley Park. These results suggest that it is unlikely to see a new graffiti location outside of the West Hastings/Dunsmuir hotspot.

Manhattan: The Manhattan graffiti dataset reports 3946 graffiti occurring at 3406 unique locations. The graffiti locations and counts are illustrated in Figure 1c. We employ the same inference strategy used for the downtown Vancouver data in Section 5. A much larger value of α_0 was required for Manhattan, which is sensible as the ratio of unique graffiti locations to total points is much higher for Manhattan than for Vancouver. Figure 1d illustrates the predictive distribution for Graffiti in Manhattan.

In this case, the posterior mean of α_0 was 11881.9, resulting in 75.1 percent of the predictive density being for new graffiti locations. Consequently, 75.1 percent of the density is assigned to G_0, shown by the contours in Figure 1d. This is a large contrast between Manhattan and Vancouver. The predictive distribution reveals high density radiating from Chinatown and into East Village. The density is also high in Harlem and down either side of Central Park. The density lowers in West Village, Chelsea, and Hell's Kitchen as well as the financial district, Columbus Circle, and corporate midtown. These findings suggest that, relative to Vancouver, the graffiti locations are much more varied, with the probability of new locations occurring being much higher.

6. Model Comparisons

In this section, we assess the quantitative performance of the ASP framework by comparing its performance to that of three different competitors: an empirical grid-based method, a DPM, and a mixture of a DPM and the empirical distribution. We describe the three competitors, provide a metric for comparison of the approaches, and illustrate the predictive performance of all models across both the downtown Vancouver and Manhattan datasets. We also demonstrate the computational expense of fitting each model. The models shown here are exemplars chosen from a range of possible competitors. Results for the larger group of competitors are available in the supplementary material.

The empirical grid-based approach is a commonly used strategy for sidestepping atomicity by dividing the area of interest into a grid with cell lengths ϵ. The density assigned to each cell is proportional to the frequency of points within each cell. The result is a piecewise continuous predictive distribution which resembles the PP model with piecewise uniform intensity proposed in Ding et al. (2012)).

The DPM we consider here is a standard Dirichlet Process mixture model of bivariate normal distributions. Despite the fact that DPMs are unable to theoretically model atomic data, we chose to include it as a competitor to demonstrate the advantage of modelling atoms explicitly.

The form of the predictive distribution for the DPM-empirical mixture is the same as that of ASP; it combines atoms at observed locations with a DPM assigning density to new locations. The two models differ in how they assign weights to these two components. In the DPM-empirical mixture, proportions are chosen using cross validation, whereas ASP uses a Bayesian resampling scheme.

The posterior predictive distribution for each model (except the empirical grid) was determined using MCMC sampling strategies like the one given in Section 4. Since the probability of observing atomic data in either DPM or empirical methods is zero almost-surely, it is unfair to compare these methods using exact locations. Instead, for each given point, we compare the predictive log likelihood of a square ϵ-region. We illustrate the results for each of the models using variety of values of ϵ. To compare across the Vancouver and Manhattan datasets, the graffiti locations are scaled to a $[-1, 1] \times [-1, 1]$ region so that an ϵ-neighborhood represents $\left(\frac{\epsilon^2}{4}\right)\%$ of the total area observed.

Performance is judged through a held-out analysis of 10% of the observed pieces of graffiti, conducting posterior inference using the reduced data, and calculating predictive likelihoods of the omitted data. For the DPM-empirical mixture, an additional 10% of the training data is held-out to cross validate the mixing proportions. In all cases, we

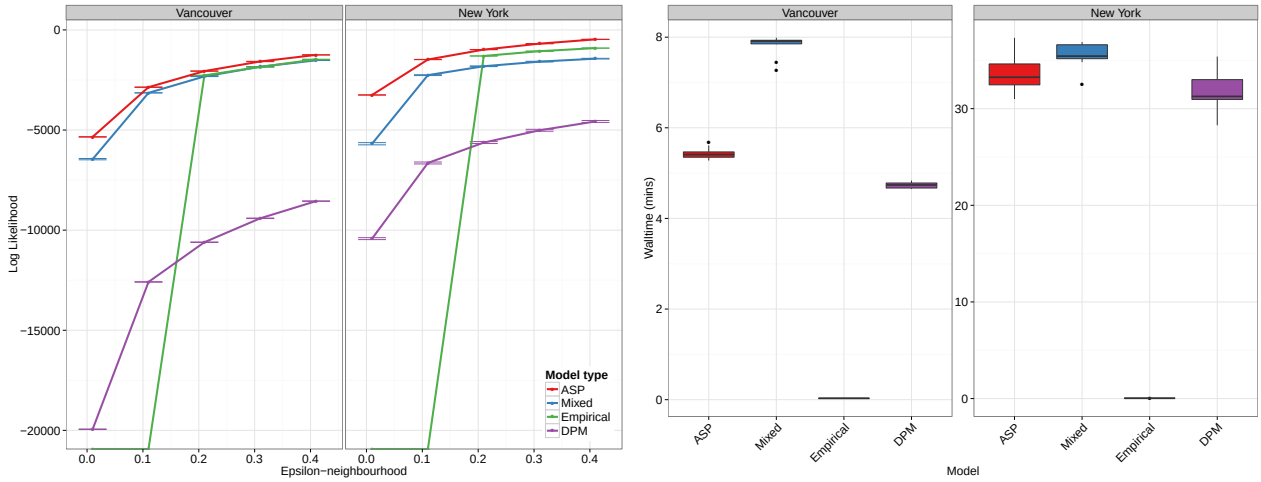

Figure 3. (Left) Predictive log-likelihood for ϵ-neighborhoods of held out pieces of graffiti for different spatial models. The ASP model uniformly outperforms the Dirichlet process mixture models (DPM), an empirical grid based approach, and a mixture between a DPM and the empirical distribution. The derivation of the predictive log-likelihood for an ϵ-neighborhood of held out pieces of graffiti is shown in the supplement. (Right) Walltime in minutes for each of the five models. The ASP model is slightly more expensive than the DPM model, but cheaper than the DPM-empirical mixture. The number of unique locations in the Manhattan dataset is almost twice the number in the Vancouver dataset. Since the Gibbs sampler is initialized to a completely disconnected configuration, the computational complexity before burn-in scales as $O(n^2)$ where n is the number of unique locations. This explains the large differences in compute time between the Vancouver and Manhattan datasets. This could be addressed using a split-merge move, which does not generally require this type of expensive initialization (Jain & Neal, 2004).

approximate the joint held-out predictive distribution with the product of the one point predictive distributions. This process is repeated for 10 separate held-out datasets. Justification of this approach is provided in the supplement.

The results of the held out analysis are shown in Figure 3. Error bars are included to show the variability between the 10 separate held-out analyses. The ASP framework predicts better than all competitors for all values of ϵ. Although the empirical grid is computationally fast and performs well for large values of epsilon, it fails for small values of epsilon due to its inability to capture the atomicity of the dataset. ASP, at worst, requires minimal computation time over DPM model while exhibiting much better predictive results. ASP also requires less time than the DPM-empirical mixture due to computational intensiveness of cross-validation. The proportions assigned to the DPM by the DPM-empirical mixture were 0.1 for Vancouver and 0.751 for Manhattan (versus 0.102 and 0.79 for ASP). Since the two approaches yield similar proportions, cross validation is not worth the computational expense and loss of held-out data.

7. Conclusion

This paper proposed an atomic spatial process (ASP) framework to model atomic urban data. A key strength of the framework is a posterior predictive distribution consisting of a mixture of discrete atoms (at previously observed locations) and a continuous measure (for new locations). This captures an important property of atomic urban data; previously observed locations have positive probability, but new locations are still possible. In contrast, a fully discrete predictive distribution allows no possibility of new locations, and a continuous predictive distribution (such as a DPM) assigns zero probability to already observed points. Our experiments further show that both DPM and empirical approaches are outperformed by ASP methods.

The novelty of our work lies not in the individual levels of the model, but in the particular nested configuration for the ASP, in the derivation of an inference method, and as a case study of BNP in spatial statistics and its evaluation. Most importantly, we can handle a spatial data type for which we have found no satisfactory models in the literature.

For future work, a dependent time aspect could be added to the modelling framework, providing a spatial-temporal version of the ASP. Another option is to consider more alternative non-parametric base measures for the GaP prior on μ. Dependent Dirichlet Processes (MacEachern, 1999) or spatial normalized random measures (Rao & Teh, 2009) could provide additional flexibility to capture structure a large heterogeneous area, such as a municipalities containing both dense urban hubs and sprawling suburbs.

Acknowledgments

This work was partially funded by an NSERC Discovery Grant and two NSERC Graduate Scholarships. Computing was supported by WestGrid. We would also like to thank Professor James Zidek and the anonymous reviewers for providing helpful comments.

References

Aldous, David J. *Exchangeability and Related Topics*. Springer, 1985.

Antoniak, Charles E. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, pp. 1152–1174, 1974.

Boruff, Bryan J, Nathan, Andrea, and Nijënstein, Sandra. Using GPS Technology to (Re)-Examine Operational Definitions of 'Neighbourhood' in Place-Based Health Research. *International Journal of Health Geographics*, 11(1):22, 2012.

Ding, Mingtao, He, Lihan, Dunson, David, and Carin, Lawrence. Nonparametric Bayesian Segmentation of a Multivariate Inhomogeneous Space-Time Poisson Process. *Bayesian Analysis (Online)*, 7(4):813, 2012.

Ferguson, Thomas S. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, pp. 209–230, 1973.

Gelfand, Alan E, Kottas, Athanasios, and MacEachern, Steven N. Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, 2005.

Gelman, Andrew, Carlin, John B, Stern, Hal S, Dunson, David B, Vehtari, Aki, and Rubin, Donald B. *Bayesian Data Analysis*. CRC Press, 2013.

Guindani, Michele, Müller, Peter, and Zhang, Song. A Bayesian Discovery Procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):905–925, 2009.

Hjort, Nils Lid, Holmes, Chris, Müller, Peter, and Walker, Stephen G (eds.). *Bayesian Nonparametrics*. Cambridge University Press, 2010.

Jain, Sonia and Neal, Radford M. A Split-merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. *Journal of Computational and Graphical Statistics*, 13(1), 2004.

Jordan, Michael I. Hierarchical Models, Nested Models and Completely Random Measures. *Frontiers of Statistical Decision Making and Bayesian Analysis: in Honor of James O. Berger. New York: Springer*, 2010.

Kingman, John Frank Charles. *Poisson Processes*, volume 3. Oxford University Press, 1992.

MacEachern, Steven N. Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55, 1999.

Malleson, Nick and Andresen, Martin A. The Impact of Using Social Media Data in Crime Rate Calculations: Shifting Hot Spots and Changing Spatial Patterns. *Cartography and Geographic Information Science*, pp. 1–10, 2014.

Mohler, George O, Short, Martin B, Brantingham, P Jeffrey, Schoenberg, Frederic Paik, and Tita, George E. Self-exciting Point Process Modeling of Crime. *Journal of the American Statistical Association*, 106(493), 2011.

Páez, Antonio and Scott, Darren M. Spatial Statistics for Urban Analysis: A Review of Techniques with Examples. *GeoJournal*, 61(1):53–67, 2005.

Rao, Vinayak and Teh, Yee W. Spatial Normalized Gamma Processes. In *Advances in Neural Information Processing Systems*, pp. 1554–1562, 2009.

Sethuraman, Jayaram. A Constructive Definition of Dirichlet Priors. Technical report, DTIC Document, 1991.

Short, Martin B, D'Orsogna, Maria R, Pasour, Virginia B, Tita, George E, Brantingham, Paul J, Bertozzi, Andrea L, and Chayes, Lincoln B. A Statistical Model of Criminal Behavior. *Mathematical Models and Methods in Applied Sciences*, 18(supp01):1249–1267, 2008.

Taddy, Matthew A. Autoregressive Mixture Models for Dynamic Spatial Poisson Processes: Application to Tracking Intensity of Violent Crime. *Journal of the American Statistical Association*, 105(492):1403–1417, 2010.

Wang, Tong, Rudin, Cynthia, Wagner, Daniel, and Sevieri, Rich. Learning to Detect Patterns of Crime. In *Machine Learning and Knowledge Discovery in Databases*, pp. 515–530. Springer, 2013.

Wang, Xiaofeng, Gerber, Matthew S, and Brown, Donald E. Automatic Crime Prediction using Events Extracted from Twitter Posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pp. 231–238. Springer, 2012.

Wolpert, Robert L and Ickstadt, Katja. Poisson/Gamma Random Field Models for Spatial Statistics. *Biometrika*, 85(2):251–267, 1998.